Basic Phylogenetic Combinatorics

Phylogenetic combinatorics is a branch of discrete applied mathematics concerned with the combinatorial description and analysis of phylogenetic trees and related mathematical structures such as phylogenetic networks and tight spans. Based on a natural conceptual framework, the book focuses on the interrelationship between the principal options for encoding phylogenetic trees: split systems, quartet systems, and metrics. Such encodings provide useful options for analyzing and dealing with phylogenetic trees and networks, and are at the basis of much of phylogenetic data processing. The book highlights how each one provides a unique perspective for viewing and perceiving the combinatorial structure of a phylogenetic tree and is, simultaneously, a rich source for combinatorial analysis and theory building. It is dedicated to Manfred Eigen who inspired many of the results presented in this book.

Graduate students and researchers in mathematics and computer science will enjoy exploring this fascinating new area, and learn how mathematics may be used to help solve topical problems arising in evolutionary biology.

ANDREAS DRESS works currently as a scientific advisor at infinity^3 GmbH, Bielefeld, Germany.

KATHARINA T. HUBER is a Lecturer in the School of Computing Sciences at the University of East Anglia, UK.

JACOBUS KOOLEN is an Associate Professor in the Department of Mathematics at Pohang University of Science and Technology (POSTECH), South Korea.

VINCENT MOULTON is a Professor in the School of Computing Sciences at the University of East Anglia, UK.

ANDREAS SPILLNER is an Assistant Professor in the Department of Mathematics and Computer Science at the University of Greifswald, Germany.

Basic Phylogenetic Combinatorics

ANDREAS DRESS (德乐思)
infinity^3 GmbH, Bielefeld, Germany
and
CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences

KATHARINA T. HUBER
University of East Anglia

JACOBUS KOOLEN
Pohang University of Science and Technology (POSTECH), South Korea

VINCENT MOULTON
University of East Anglia

ANDREAS SPILLNER
University of Greifswald

CAMBRIDGE UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521768320

First published 2012

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Basic phylogenetic combinatorics / Andreas Dress . . . [et al.].
p. cm.
ISBN 978-0-521-76832-0 (Hardback)
1. Branching processes. 2. Combinatorial analysis. I. Dress, Andreas.
QA274.76.B37 2011
511´.6–dc23

2011043264

ISBN 978-0-521-76832-0 Hardback

We dedicate this book to Manfred Eigen whose
questions concerning the evolution of RNA and DNA sequences
inspired many of the early results in
phylogenetic combinatorics that ultimately led
to the work presented in this book.

Contents

Preface

More than one and a half centuries have passed since Charles Darwin presented his theory on the origin of species asserting that all organisms are related to each other by common descent via a "tree of life". Since then, biologists have been able to piece together a great deal of information concerning this tree — relying in particular in more recent times on the advent of ever cheaper and faster DNA sequencing technologies. Even so, there remain many fascinating open problems concerning the tree of life and the evolutionary processes underlying it, problems that often require sophisticated techniques from areas such as mathematics, computer science, and statistics.

Phylogenetic combinatorics can be regarded as a branch of discrete applied mathematics concerned with the combinatorial description and analysis of *phylogenetic* or *evolutionary trees* and related mathematical structures such as phylogenetic networks, complexes, and tight spans. In this book, we present a *systematic* approach to phylogenetic combinatorics based on a natural conceptual framework that, simultaneously, allows and forces us to encompass many classical as well as a good number of new pertinent results.

More specifically, this book concentrates on the interrelationship between the three principal ways commonly used for **encoding** phylogenetic trees: *Split systems*, *metrics*, and *quartet systems* (see Figure 1). Informally, for X some finite set, a split system over X is a collection of bipartitions of X, a quartet system is a collection of two-versus-two bipartitions of subsets of X of size four, and a metric is a bivariate function assigning a "distance" to any pair of elements in X.

Such encodings provide useful options for analyzing and manipulating phylogenetic trees with leaves labeled by X, and are at the basis of much of phylogenetic data processing. Indeed, they arise naturally from the various types of data from which phylogenetic trees are typically (re-)constructed: Comparative sequence analysis of genes or genomes may lead to metrics, character tables as

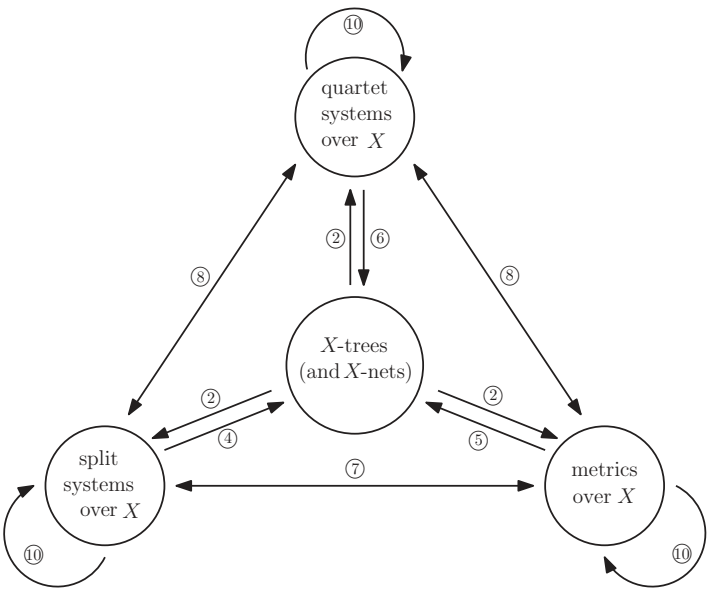

Figure 1 In this figure, we indicate the manifold relationships between various combinatorial objects relevant in phylogenetic analysis that will be studied in this book.

well as single nucleotide polymorphisms give rise to splits (and metrics), and quartet systems arise from restricting attention in phylogenetic data analysis to just four taxa at a time to avoid reconstruction algorithms becoming overwhelmed by the sheer number of taxa that need to be treated simultaneously. All three types of encodings require a solid theoretical foundation and provide, at the same time, a rich source for combinatorial analysis and theory building.

This book aims to highlight how each of the three types of encodings provides a unique perspective for viewing the combinatorial structure of a phylogenetic tree, for assessing the suitability of given data for tree reconstruction, and, if suitable, for recovering such trees from such data. And it will, of course, also discuss how split systems, metrics, and quartet systems are related to one another.

Here is an outline of the contents: After presenting some basic definitions and concepts that will be used throughout the book in Chapter 1, we introduce the formal concept of a phylogenetic tree or, a bit more generally, an X-tree in Chapter 2. We then define split systems, metrics, and quartet systems, and show that X-trees may indeed be uniquely encoded in terms of such combinatorial objects. In Chapter 3, we then proceed to identify which split systems, metrics,

or quartet systems are induced by — and thus encode — an X-tree: That is, we characterize the split systems, metrics, or quartet systems in the "image" of the "maps" labeled ② in Figure 1 in terms of some simple, yet instructive and enlightening conditions.

In Chapters 4, 5, and 6, we move on to the problem of deciding how to *decode* a given tree-encoding using appropriate constructions corresponding, respectively, to the maps labeled ④, ⑤, ⑥ in Figure 1. In other words, given a split system, a metric, or a quartet system in the image of the maps labeled ②, we consider how to find that (unique!) X-tree encoded by them. We will also explain how, when applied to data sets that do not encode a tree, these constructions can produce *networks* (as opposed to trees) and discuss some pertinent consequences.

In Chapters 7 and 8, we investigate the *recoding* problem: How can we compute the split system, metric, or quartet system encoding a tree from its other two encodings, i.e., how can we define maps (as indicated by the arrows labeled ⑦ and ⑧, in either direction) so that the resulting triangular subdiagrams in Figure 1 are commutative? Generally, the pertinent constructions lead to pairs of maps between any two of the three classes of objects that appear to be of some independent interest in themselves.

"Rooting" an X-tree is rather important for a realistic interpretation of X-trees in terms of evolutionary history. Correspondingly, we consider in Chapter 9 how the previously mentioned maps and constructions can be modified so as to apply to *rooted X-trees*, giving rise to *cluster systems* (rather than split systems), *triplet systems* (rather than quartet systems), and *hierarchical dissimilarities* and *ultrametrics* (rather than metrics). Mathematically speaking, this can be regarded as taking an *affine* (more concrete) versus a *projective* (more elegant) approach to working with phylogenetic trees.

In the final chapter, Chapter 10, we address the problem of how to measure and remove "inconsistencies" in split systems, metrics, and quartet systems. In other words, given one such structure that does not encode an X-tree, we explore how we may find another one in its "neighborhood" that does. As we shall see, this not only yields some interesting mathematical results, but also new ways to analyze and understand phylogenetic data.

A major feature of this book is that full proofs are provided for all of the fundamental results, thus giving the motivated reader a chance to get to the forefront of the field of phylogenetic combinatorics without having to spend too much time seeking references (see also Figure 2 for a *Leitfaden*, i.e., an overview of chapter dependencies). It also includes various new results and proofs that have not been published previously, and it attempts to introduce most topics in an elementary way. Overall, we hope that the reader will be

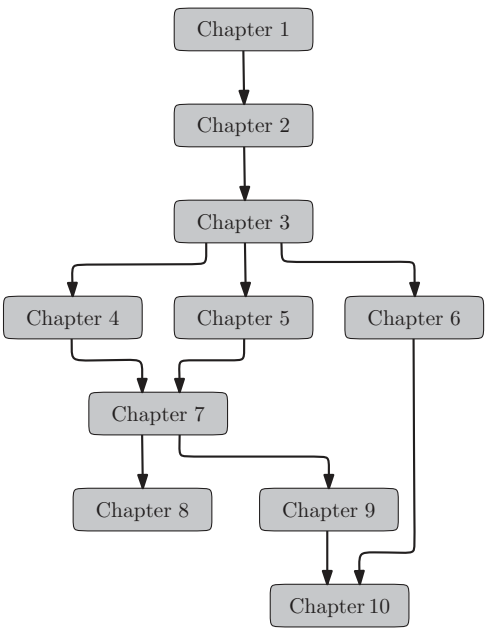

Figure 2 A diagram depicting the main dependencies between chapters in this book.

motivated by the book to explore this interesting area of mathematics whilst, at the same time, having the opportunity of seeing how mathematics may be used to help with solving topical problems that arise outside mathematics.

Finally, we would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for giving us the opportunity to jointly draft and work on this book there, and also the UK Engineering and Physical Sciences Research Council and Royal Society, the Basic Science Research Program through the National Research Foundation of Korea (NRF) (grant number 2010-0008138), the DFG and the Max Planck Society, Germany, and the Chinese Academy for Sciences for financial support. We also thank our friends and colleagues and, in particular, David Bryant, Stefan Grünewald, Mike Hendy, Daniel Huson, Saitou Naruya, David Penny, LI Qiang, Charles Semple, Mike Steel, and WU Yaokun for many stimulating, critical, as well as encouraging discussions and comments. In addition, we thank students and colleagues at Bangalore, Bandar Lampung, Christchurch, Manila, New York, Paris, Pohang, and Shanghai, for their helpful feedback in courses where early versions of this text were presented. And last but not least, we all thank our families and, especially, Christiana, Keiko, Therese, Eugen, Jacky, and Robin, for their patience and support.